Asynchronously Embedding Psychological Test Questions into Human–Robot Conversations for User Profiling

Tsung-Ren Huang^{1,5,6} · Yu-Wei Liu^{1,5} · Shin-Min Hsu^{1,5} · Joshua O. S. Goh^{2,5} · Yu-Ling Chang^{1,5} · Su-Ling Yeh^{1,5} · Li-Chen Fu^{3,4,5}

Accepted: 24 October 2020 / Published online: 20 November 2020
© The Author(s) 2020

Abstract

Psychological variables of a person (e.g., cognitive abilities, personality traits, emotional states, and preferences) are valuable information that can be utilized by social robots to offer personalized human–robot interaction. These variables are often latent and inferred indirectly from a third-person perspective based on an individual’s behavioral manifestations (e.g., facial emotion expressions), and hence the true values of inferred psychological variables remain unknown to a robot observer. Although earlier studies have employed robot-administered psychological tests to infer psychological variables based on an individual’s first-person responses, these tests were formally presented and could be tedious to some users. To leverage the validity and reliability of well-established psychological tests for user profiling with ease, the present study examined the possibility of asynchronously embedding psychological test questions into casual human–robot conversations. In our experiment using a big-five personality inventory, the verbal responses from users to these asynchronous test questions were then compared with the written responses to the same personality test. The personality measures estimated from the two approaches correlated strongly in a young adult population but only moderately in an older population. These findings demonstrate the validity of the proposed asynchronous method for psychological testing in human–agent interactions and suggest some caveats when this testing method is applied to older adults or other special populations.

Keywords Human-centered robotics · Human–robot interaction · Psychological testing

1 Introduction

Social robots can take many different forms [1, 2] and the ones that interact with humans—companions or helpers—have become increasingly demanded [3]. For example, socially interactive robots [4] have moved from research lab-

oratories into shopping malls [5] and consumer markets (e.g., Cozmo & RoBoHoN) for providing service and entertainment. According to the International Federation of Robotics, 1.7 million social and entertainment robots were sold in 2015, and the number of sales is projected to be over 7.4 million units in 2025 [6]. In another example, socially assistive robots [5] have been gradually adopted to assist in the health care of children [7], adults [8], and the elderly [9].

For successful long-term interactions with humans, social robots are recommended to identify returning users, remember past interactions with them, and get to know them for offering personalized responses [10, 11]. For example, a user may be annoyed by the same, non-personalized greeting question from a robot: “*Do you mind telling me more about yourself?*” For another example, the personal spatial zone a user prefers during human–robot interaction is influenced by the personality of that user [12]. Therefore, it is important for a social robot to learn the personal information and idiosyncrasies of a user so as to foresee and meet the social, emotional, and cognitive needs of that user [13].

✉ Tsung-Ren Huang
tren@mil.psy.ntu.edu.tw

¹ Department of Psychology, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd, Taipei 10617, Taiwan

² Graduate Institute of Brain and Mind Sciences, National Taiwan University, Taipei, Taiwan

³ Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

⁴ Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

⁵ MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan

⁶ MOST AI Biomedical Research Center, Tainan, Taiwan

Despite its importance, understanding or profiling a user is a challenging task, especially when concerning hidden psychological variables (e.g., traits, states, values, and preferences). Although not directly observable, a user's psychological variables (e.g., personality) has been inferred indirectly from behavioral manifestations (e.g., linguistic or prosodic patterns [14]) or psychological tests (e.g., a personality questionnaire [15]) to facilitate human–robot interaction (HRI). Note, however, that inferences from behavioral manifestations usually rely on observations from a *third-person* perspective, and hence the true values of inferred psychological variables actually remain unknown to an observer, be it a social robot or a real person. By contrast, psychological tests, particularly those with sound psychometric properties, can reliably estimate the values of an individual's psychological variables based on the individual's *first-person* responses.

Although psychological tests have been utilized in previous HRI studies, overall they have not been heavily employed as user-profiling tools and used to their full potential for improving HRI. In most HRI studies, psychological tests were administered by researchers using paper questionnaires to understand a user before or after HRI [15] rather than during HRI for a robot to dynamically adapt its behavior to individual users. Only a few recent studies pioneered the possibility of having a robot to administer psychological tests during HRI, and all of these robot-administered psychometric evaluations used standard testing procedures as in human-administered tests (e.g., [16]), which may sometimes appear uninteresting or even tedious to users.

To circumvent the formality problem, here we propose using asynchronous test questions (ATQs) as a user-friendly way of administering a psychological test during HRI. Specifically, our proposed testing procedure consists of three steps: (1) obtaining items or questions of a psychological test; (2) embedding parts of these questions into contextually relevant periods in a conversation during HRI as asynchronous mini-tests; and (3) aggregating all the answers to these ATQs for scoring the psychological test. Compared to a traditional psychological test that usually assesses one psychological domain with items temporally grouped together, ATQs from a psychological test are not given to a testee all at once. As a result, ATQs that assess a psychological domain as a whole are less susceptible to the issues of sustained attention and cognitive demand relative to their original test. Furthermore, because ATQs are presented casually in contextually relevant periods in a conversation rather than formally in a test setting, it is less likely for a testee to be self-conscious about being tested and modify his/her responses accordingly [17].

As has been reported previously, different formats of the same psychological assessment—such as computer, questionnaire, or interview—could lead to differences in evaluations [18]. To examine whether a temporally fragmented psychological test could yield results comparable to those

of the original test and thus substitute for the original test, the present study used a social robot to administer a big-five personality test amidst a broader HRI session for each research participant using the ATQ procedure because big-five personality measures are commonly used predictors of human behavior [19, 20] and important dimensions in HRI [21]. The verbal responses from each participant to ATQs were then compared with the written responses to the same personality test for validating the ATQ procedure. The ideal result would be a perfect positive correlation between verbal and written responses to the same test, although theoretically the upper bound of this correlation is the test-retest reliability of the psychological test. The procedure and results of our ATQ experiment are detailed in the following sections.

2 Methods

The ATQ experiment was part of a larger human–robot interaction study on both young and older adults. Approved by the Research Ethics Office of National Taiwan University (REC 201803HS017), the 1-h larger study consisted of three main HRI events. They are robot-administered cognitive testing, followed by robot-accompanied toy-playing, and then robot-assisted tablet-using. All participants gave written informed consent for their participation in the study.

2.1 Human Participants

Relative to young adults, older adults tend to have stronger negative attitudes toward robots [22], which may, in turn, affect their human–robot interactions in general and our robot-administered psychological testing in particular. Therefore, we experimented with our ATQ procedure on two age groups of participants to examine its suitability for use with both young and older adults, which cannot be taken for granted [23].

There were 26 participants in the young group (13 males and 13 females; mean age: 21.42; age range: 18–29) and 20 participants in the older group (11 males and 9 females; mean age: 72.25; age range 67–81). None of these participants had auditory impairments. Each experimental session consisted of a participant interacting with a social robot in a well-lit room with a one-way mirror, with the whole HRI session recorded by a hidden camera.

2.2 Social Robot

We used a programmable humanoid robot—RoBoHoN (Sharp Co., Ltd.)—in our study. RoBoHoN is 19.5 cm tall when standing. For research participants to maintain eye contact with RoBoHoN during HRI, the RoBoHoN unit used in our experiment was placed on a table to converse with the

Fig. 1 A side view of the HRI environment

seated research participant (Fig. 1). RoBoHoN has a built-in speech-to-text and text-to-speech engines for speech recognition and production, respectively. Although it could be fully autonomous, our RoBoHoN was remotely controlled by a human operator in a dark observation room to accurately detect sentence endpoints in participant speech and manage relevant conversational contingencies such as speech pauses, repetitions, queries, etc. that arise during memory recall and decision-making.

2.3 Psychological Test

We used the Ten Item Personality Inventory (TIPI) in that each of the five personality dimensions—*Extraversion*, *Agreeableness*, *Conscientiousness*, *Emotional Stability*, & *Openness to Experience*—is measured by two items [24]. The personality inventory was administered twice on each participant to validate whether the original test can be substituted by its ATQs. The first test was administered verbally by RoBoHoN using ATQs across each 1-h HRI session (Fig. 2), whereas the second test was administered by an experimenter right after the whole HRI session as a post-study questionnaire in TIPI's original form on paper. The order of the two administrations was not counterbalanced to prevent participants from experiencing ATQs as repeated “test” questions, which would defeat the purpose of using ATQs in non-testing contexts. Both administrations imposed a 5-point response format on each item: *Strongly Agree* = 5, *Slightly Agree* = 4, *Neutral* = 3, *Slightly Disagree* = 2, and *Strongly Disagree* = 1. Specifically, RoBoHoN would ask: “Do you strongly or slightly agree, strongly or slightly disagree, or are you neutral?” as a follow-up question to verbally described these five response options whenever a participant's response to an ATQ could not be clearly mapped onto the 5-point Likert scale.

Fig. 2 Normalized temporal positions of 10 ATQs in each 1-h HRI session. Each color represents a different ATQ to show the variability across sessions. Each number associated with each color corresponds to the item number listed in Tables 1 and 2, and zero indicates other HRI events that are outside the scope of the present study

The ATQs used in this study were not a verbatim copy of the original TIPI test [24]. They were adapted to appear natural in human–robot conversations. The ten pairs of personality-describing adjectives, as shown in Tables 1 and 2, were embedded in sentences such as “Are you, in general, an [anxious, easily upset] person?”, “Do you consider yourself a [conventional, uncreative] person in general?” or “I’m curious whether you are a [dependable, self-disciplined] person in general.” The original instruction of TIPI is: “Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.”

It should be noted that the human–robot conversational contexts were prearranged such that the 10 pre-programmed ATQs were asked by the RoBoHoN as parts of small talk before or after an interaction event, such as toy-playing or cognitive testing. As a result, the presentation order of the 10 personality items was fixed across participants and not identical to the order in the original test. For example, when RoBoHoN greeted a participant at the beginning of the study, it expressed interest in learning more about the participant and asked whether he/she is a conventional person. In another example, right before a toy-playing event, RoBoHoN asked a participant whether he/she is a person who is open to new experiences.

3 Results

For the comparison between a participant's verbal and written responses to the same psychological test, the participants'

verbal responses to ATQs were initially coded from video recordings by two independent coders from our research team who knew the purpose of the present study but were not provided with the participants' written responses at the time of behavioral coding. The coding instructions were as follows. "Please help label the degree of agreement expressed in each participant's verbal response by a number from one to five with one being a strong disagreement, two being a slight disagreement, three being neutral, four being a slight agreement, and five being a strong agreement. If a particular response of a participant is difficult to judge, please make your best guess based on the response patterns of that participant, if any."

There were indeed difficult cases and minor coding differences (ordinal Krippendorff' $\alpha = 0.995$ and 0.87 for the younger and older groups, respectively). For example, an older participant shook her head without a verbal response to an ATQ but then verbally answered "Agree" when RoBo-HoN reminded her of the five response options. The two raters discussed in person on their coding differences incurred by such difficult cases to reach a consensus on each item score of each verbal response.

The final item scores were then compared with the scores obtained from the written responses of the same participants. Below we report the results of Pearson correlations, paired t-tests, and their associated equivalence tests [25]. For each test item or personality dimension, an ideal result would be a perfect positive correlation and even no difference between verbal and written responses, which translates to a correlation not statistically equivalent to 0 and significantly larger than 0 as well as a difference not significantly different from 0 and statistically equivalent to 0.

3.1 The Young Adult Group

The results of the younger participants are summarized in Table 1. For almost all the ten test items and five personality dimensions, both null-hypothesis and equivalence tests suggest very strong positive correlations and little differences between the participants' verbal and written responses. In other words, the young participants responded similarly to the same questions regardless of whether these questions were synchronously or asynchronously presented, suggesting a good validity of our proposed ATQ testing procedure on the younger adult population.

3.2 The Older Adult Group

The analysis results of the older participants are summarized in Table 2. Overall, the results were not as ideal as the ones of the young participants. Several correlations between the two response formats were weak or even statistically not different from 0. Moreover, the distributions of verbal and written responses are not statistically equivalent for some

test items and personality dimensions. In particular, the ratings of one's *Openness to Experiences* trait were not very consistent between the two testing procedures, suggesting some potential problems in the use of the ATQ procedure on the older adult population.

4 Discussion

We used a brief personality test as an example to explore the possibility of using temporally fragmented psychological tests for user-friendly psychological assessment in human–robot interaction. As the first attempt toward such endeavor, the present study examined whether a psychological test, when administered informally in the form of asynchronous test questions, could still lead to assessment results comparable to those from a formal psychological test. Our results showed that the proposed ATQ testing procedure was quite successful with younger adults but only partially successful with older adults.

The discrepancies between participants' verbal and written responses, especially those observed in older adults, could result from several sources. First, participants, particularly those low in behavioral consistency, might not respond consistently to the same test question even with the same administration method. Second, participants, when receiving the same question through different perceptual modalities (audition vs. vision), might engage distinct attentive/memory processes [26, 27] and thus make different choice decisions. Third, participants might mishear some questions unnaturally pronounced by a robot, a situation that had been observed during robot-administered cognitive assessments [16]. Fourth, participants might be less aware of being tested and hence less defensive when answering informal ATQs than formal test questions. Fifth, when asked the same question, participants might be more willing to disclose themselves to a robot than a researcher. Previous studies have found people to be more willing to share personal information with non-judgmental computer agents [18, 28].

While the aforementioned factors might all affect the experimental results and the current study design could not distinguish these different sources of contributions, the proposed ATQ procedure, when applied to the young adults, still yielded results on par with those from a conventional paper-based test. By contrast, some of these factors might affect older adults more pronouncedly than younger adults. Revisiting the video recording of each HRI session allowed us to provide some speculations as to why the ATQ procedure did not work as well in the older adults and how this might be overcome in the future. Below we will elaborate on these findings together with other implications from this explorative study.

Table 1 The correlations and differences in means between verbal and written responses of the young group (N = 26)

	Type of responses (mean $\pm$ sd)		Pearson correlation		Paired <i>T</i> test	
	Verbal	Written	<i>r</i>	90% TOST CI	<i>t</i> _{null}	<i>t</i> _{eq}
<i>Traits</i>						
1. Extraversion	3.12 $\pm$ 1.03	3.12 $\pm$ 0.93	0.88***	[0.78, 0.94]	0.00	− 2.91**
2. Agreeableness	3.75 $\pm$ 0.85	3.63 $\pm$ 0.87	0.83***	[0.69, 0.91]	1.19	− 1.72*
3. Conscientiousness	3.77 $\pm$ 0.92	3.62 $\pm$ 0.93	0.77***	[0.59, 0.88]	1.25	− 1.66
4. Emotional stability	3.21 $\pm$ 0.98	3.17 $\pm$ 0.92	0.87***	[0.76, 0.93]	0.40	− 2.51**
5. Openness to experience	3.85 $\pm$ 0.86	3.79 $\pm$ 0.84	0.77***	[0.60, 0.88]	0.52	− 2.40*
<i>Items</i>						
1. Critical, quarrelsome	2.65 $\pm$ 1.35	2.81 $\pm$ 1.41	0.76***	[0.57, 0.87]	− 0.81	2.10*
2. Disorganized, careless	2.31 $\pm$ 1.32	2.54 $\pm$ 1.21	0.67***	[0.44, 0.82]	− 1.14	1.77*
3. Conventional, uncreative	2.27 $\pm$ 1.19	2.31 $\pm$ 1.12	0.81***	[0.65, 0.90]	− 0.27	2.63**
4. Anxious, easily upset	3.38 $\pm$ 1.30	3.31 $\pm$ 1.26	0.86***	[0.73, 0.92]	0.57	− 2.34*
5. Extraverted, enthusiastic	3.27 $\pm$ 1.22	3.38 $\pm$ 1.13	0.85***	[0.72, 0.92]	− 0.90	2.01*
6. Open to new experiences, complex	3.96 $\pm$ 1.04	3.88 $\pm$ 1.03	0.82***	[0.67, 0.90]	0.63	− 2.28*
7. Sympathetic, warm	4.15 $\pm$ 0.83	4.08 $\pm$ 0.63	0.82***	[0.67, 0.90]	0.81	− 2.10*
8. Reserved, quiet	3.04 $\pm$ 1.15	3.15 $\pm$ 0.97	0.79***	[0.62, 0.89]	− 0.83	2.08*
9. Calm, emotionally stable	3.81 $\pm$ 1.13	3.65 $\pm$ 0.94	0.84***	[0.71, 0.92]	1.28	− 1.63
10. Dependable, self-disciplined	3.85 $\pm$ 1.05	3.77 $\pm$ 0.95	0.85***	[0.72, 0.92]	0.70	− 2.21*

For Pearson correlations, the 90% confidence intervals (CI) of the two one-sided tests (TOST) were compared to the zero-equivalent range = [− 0.39, 0.39] with power set to be 0.8, and none of the Pearson correlations shown in this table were statistically equivalent to 0. For the two-tailed paired *T* tests, *t*_{null} were *t* values of the null-hypothesis tests, and none of them suggested a difference in means between verbal and written responses; *t*_{eq} were *t* values of the zero-equivalence tests, most of which suggested the differences in means to be statistically equivalent to 0 when power set to be 0.8 and Cohen's *d*_z = 0.57

P* < 0.05; *P* < 0.01; ****P* < 0.001

4.1 Validity, Reliability, and Applicability of ATQs

The validity and reliability of ATQs are limited by the ones of their original psychological test. For example, in our younger group, the Pearson correlations between participants' verbal and written responses ranged from 0.77 (*Openness to Experience*) to 0.88 (*Extraversion*), which are comparable to the six-week test-retest reliability that ranged from 0.62 (*Openness to Experience*) to 0.77 (*Extraversion*) in the original TIPI study on young adults [24]. Additionally, as evidenced by the marked differences between the results of the two age groups in our study, the psychometric properties of ATQs may vary with different tested populations in a way similar to their original test, such as an overall reduced test-retest

reliability of TIPI in older adults relative to younger adults [29].

The applicability of ATQs is critically constrained by the cognitive capabilities of testees. It is important to note that questions asynchronously presented in a conversation and questions simultaneously presented on a paper are processed respectively through audition and vision, two fundamentally different perceptual modalities. Thus, compared to the questions of a paper-based test that can be seen all at once, ATQs can only be heard sequentially, and a testee cannot voluntarily reinspect words that have been presented or modify an answer to an earlier ATQ. Consequently, if a testee cannot maintain auditory attention or short-term memory throughout the entire presentation of an ATQ, he/she may not process the question as fully as in the case of a paper-based test.

Table 2 The correlations and differences in means between verbal and written responses of the older group (N = 20)

	Type of responses (mean $\pm$ sd)		Pearson correlation		Paired <i>T</i> test	
	Verbal	Written	<i>r</i>	90% TOST CI	<i>t</i> _{null}	<i>t</i> _{eq}
<i>Traits</i>						
1. Extraversion	3.23 $\pm$ 1.02	3.12 $\pm$ 0.72	0.62**	[0.32, 0.81]	0.56	− 2.40*
2. Agreeableness	4.08 $\pm$ 0.69	4.12 $\pm$ 0.79	0.65**	[0.36, 0.83]	− 0.36	2.60**
3. Conscientiousness	4.47 $\pm$ 0.66	4.38 $\pm$ 0.65	0.77***	[0.54, 0.89]	1.00	− 1.95*
4. Emotional stability	3.80 $\pm$ 0.80	4.10 $\pm$ 0.77	0.61**	[0.30, 0.80]	− 1.93	1.02
5. Openness to experience	3.48 $\pm$ 1.01	3.83 $\pm$ 0.91	0.18	[− .21, 0.53]	− 1.28	1.68
<i>Items</i>						
1. Critical, quarrelsome	2.40 $\pm$ 1.23	2.30 $\pm$ 1.38	0.64**	[0.34, 0.92]	0.40	− 2.55**
2. Disorganized, careless	1.50 $\pm$ 0.69	1.65 $\pm$ 0.81	0.52*	[0.17, 0.75]	− 0.90	2.05*
3. Conventional, uncreative	2.85 $\pm$ 1.39	2.40 $\pm$ 1.05	0.37	[− .01, 0.66]	1.44	− 1.51
4. Anxious, easily upset	2.75 $\pm$ 1.29	2.05 $\pm$ 1.05	0.55*	[0.22, 0.77]	2.77*	− 0.18
5. Extraverted, enthusiastic	3.85 $\pm$ 1.14	3.80 $\pm$ 0.89	0.80***	[0.60, 0.90]	0.33	− 2.63**
6. Open to new experiences, complex	3.80 $\pm$ 0.89	4.05 $\pm$ 1.10	0.17	[− .22, 0.52]	− 0.87	2.09*
7. Sympathetic, warm	4.55 $\pm$ 0.60	4.55 $\pm$ 0.83	0.42	[0.04, 0.69]	0.00	− 2.95**
8. Reserved, quiet	3.40 $\pm$ 1.27	3.55 $\pm$ 1.23	0.29	[− .10, 0.60]	− 0.45	2.50*
9. Calm, emotionally stable	4.35 $\pm$ 0.75	4.25 $\pm$ 0.72	0.62**	[0.31, 0.81]	0.70	− 2.25*
10. Dependable, self-disciplined	4.45 $\pm$ 0.69	4.40 $\pm$ 0.82	0.78***	[0.58, 0.90]	0.44	− 2.51*

For Pearson correlations, the 90% confidence intervals (CI) of the two one-sided tests (TOST) were compared to the zero-equivalent range = [− 0.44, 0.44] with power set to be 0.8, and none of the Pearson correlations shown in this table were statistically equivalent to 0. For the two-tailed paired *T* tests, *t*_{null} were *t* values of the null-hypothesis tests, and only the one concerning "Anxious, easily upset" suggested a difference in means between verbal and written responses; *t*_{eq} were *t* values of the zero-equivalence tests, most of which suggested the differences in means to be statistically equivalent to 0 when power set to be 0.8 and Cohen's *d*_z = 0.66

p* < 0.05; *p* < 0.01; ****p* < 0.001

As a case in point, the older participants in our study might not process ATQs thoroughly. For example, when responding to the test item, "Do you consider yourself a conventional, uncreative person in general?", many of our elderly participants paused, pondered, and then responded, "Yes, I'm conventional." Such a narrow focus on parts of a heard sentence could result from an age-associated, smaller capacity of attention and short-term memory [30] or a developed habit of selective attention [31]. This can explain why older participants agreed more with the compound question "conventional, uncreative" in their verbal than written responses—sequential auditory processing might have accentuated the earlier "conventional" part [32], with which they agreed more relative to the later "uncreative" part.

Moreover, a cognitively demanding task that precedes particular ATQs may set up a stressful or even frustrating context and induce age-based stereotype threats [33, 34] in older but not younger adults when they answer those ATQs. In our study, the ATQ "anxious, easily upset" was asked right after a series of RoBoHoN-administered cognitive tests in which the older participants showed much poorer performance (not shown here) than did the younger participants, presumably because of cognitive decline [35–37]. Such results suggest that these tests might be rather challenging to our older participants and therefore induce their stress responses, including negative affect [38, 39]. By contrast, the paper version of our personality test was given right after a task relatively easy for both younger and older participants and could be the rea-

son behind why our older participants agreed less with being “anxious, easily upset” in their written than verbal responses.

4.2 Guidelines for Using ATQs

Based on the above discussions, we recommend the following for effective use of ATQs:

- (1) Basing questions on a psychological test that has good validity and reliability to avoid invalid measures, such as the unreliably measured *Openness to Experience* in the present study;
- (2) Adapting questions for the target population in a way unambiguous to them to avoid unintended responses, such as the partial answers of our older participants’ to the compound questions;
- (3) Asking the same question on different occasions if the answer to this question may lack cross-situational consistency, such as the responses of our older participants’ to the “anxious, easily upset” question.

Importantly, despite the fact that the validity and reliability of ATQs may vary across populations, it is unnecessary and impractical to also administer a paper version of the same test in addition to ATQs, especially when ATQs are applied in commercial products. This is because the test–retest reliability of ATQs can be directly evaluated using our third recommendation, and arguably the ultimate validity of ATQ is whether they help estimate internal variables of a person and improve predictions on the behaviors of that person [40, 41].

Last but not least, because estimated personal characteristics may be exploited for malicious purposes [42], such sensitive information should be securely protected [43], and users have the right to decline any form of psychological testing in the first place [44], including ATQs. One possible way to address users’ privacy concerns is to ask users’ informed consent [45] so that users are aware of being profiled. It also helps to solve the personalization-privacy paradox if personal information can be stored locally inside a conversational agent rather than in the cloud. This approach has been proved effective in reducing smartphone users’ perception of privacy violations [46].

4.3 Possible Future Directions

The present study has several limitations and can be extended in the following directions.

- Exploring the longest temporal window within which the ATQ procedure can remain valid;
- Using a psychological test with more items to improve test reliability [47, 48] ;

- Embedding a question into a contextually relevant conversation by an automatic information retrieval mechanism [49];
- Taking non-Likert, natural language responses from participants as they are and scoring them by fuzzy methods [50].

These possible extensions are of importance in applications. They will either clarify the boundary conditions or automate the question-distributing and answer-scoring components of the ATQ testing procedure. Future feasibility studies are needed to address these important issues and put ATQs into use in real-life HRI.

5 Conclusion

The present study put forward and experimented with the possibility of asynchronously administering a psychological test by embedding items from the test into human–robot conversations. This proposed ATQ procedure was then successfully validated on a young adult group but less so on an older group, based on which we derived our guidelines for future effective use of this approach.

The ATQ procedure is designed as a user-friendly method of psychological testing during human–robot interaction. Social robots can leverage such a non-strenuous procedure to support fragile individuals in the completion of psychological tests or to profile general users for response personalization. Overall, the asynchronous testing procedure holds great promise for improving user understanding and thereby human–robot connections.

As a concluding remark, it should be pointed out that the ATQ testing procedure can, in theory, be generalized to use various psychological tests and applied to various populations once the test questions are properly adapted for a target population. Also, this proposed method can be implemented in various social robots designed for companion or assistance, such as text-based chatbots [51] and embodied conversational agents [52]. All in all, we hope that the ATQ testing procedure can help import the long-accumulated knowledge of psychology—into robotics for improving machine cognition and service.

Acknowledgements The authors wish to thank Yi-Jhen Chen, Chih-Wei Ning, Yu-Lan Cheng, and Yu-Wei Lu for collecting the data of the young group, as well as Hsin-Yi Hung, Yun-Shiuan Chuang, and Uan-Luen Hsieh for collecting the data of the older group.

Funding This study was funded by the grants 109-2634-F-002-027 (under the MOST Joint Research Center for AI Technology and All Vista Healthcare) and 109-2634-F-002-023 (under the MOST AI Biomedical Research Center) from the Ministry of Science and Technology in Taiwan.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Dautenhahn K (2007) Socially intelligent robots: dimensions of human–robot interaction. *Philos Trans R Soc B* 362(1480):679–704. <https://doi.org/10.1098/rstb.2006.2004>
- Hegel F, Muhl C, Wrede B, Hielscher-Fastabend M, Sagerer G (2009) Understanding social robots. In: Proceedings of the second international conferences on advances in computer–human interactions (ACHI), Cancun, Mexico, 2009. IEEE, pp 169–174. <https://doi.org/10.1109/ACHI.2009.51>
- Korn O (2019) Social robots: technological, societal and ethical aspects of human–robot interaction. *Hum Comput Interact Ser*. <https://doi.org/10.1007/978-3-030-17107-0>
- Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robot Auton Syst* 42(3–4):143–166. [https://doi.org/10.1016/S0921-8890\(02\)00372-X](https://doi.org/10.1016/S0921-8890(02)00372-X)
- Kanda T, Shiomi M, Miyashita Z, Ishiguro H, Hagita N (2010) A communication robot in a shopping mall. *IEEE Trans Robot* 26(5):897–913. <https://doi.org/10.1109/TRO.2010.2062550>
- Murphy A (2017) Social and entertainment: robotics outlook 2025. <https://loupventures.com/social-and-entertainment-robotics-outlook-2025/>. Accessed 26 Aug 2019
- Moerman CJ, van der Heide L, Heerink M (2018) Social robots to support children's well-being under medical treatment: a systematic state-of-the-art review. *J Child Health Care* 23(4):596–612. <https://doi.org/10.1177/1367493518803031>
- Scoglio AA, Reilly ED, Gorman JA, Drebing CE (2019) Use of social robots in mental health and well-being research: systematic review. *J Med Internet Res* 21(7):e13322. <https://doi.org/10.2196/13322>
- Abdi J, Al-Hindawi A, Ng T, Vizcaychipi MP (2018) Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open* 8(2):e018815. <https://doi.org/10.1136/bmjopen-2017-018815>
- Gockley R, Bruce A, Forlizzi J, Michalowski M, Mundell A, Rosenthal S, Sellner B, Simmons R, Snipes K, Schultz AC (2006) Designing robots for long-term social interaction. *Int Conf Intell Robot Syst (IROS2006)*:1338–1343. <https://doi.org/10.1109/IROS.2005.1545303>
- Leite I, Martinho C, Paiva A (2013) Social robots for long-term interaction: a survey. *Int J Soc Robot* 5(2):291–308. <https://doi.org/10.1007/s12369-013-0178-y>
- Walters ML, Dautenhahn K, Te Boekhorst R, Koay KL, Kaouri C, Woods S, Nehaniv C, Lee D, Werry I The influence of subjects' personality traits on personal spatial zones in a human–robot interaction experiment. In: ROMAN 2005. IEEE international workshop on robot and human interactive communication, 2005., 2005. IEEE, pp 347–352
- Dautenhahn K Robots we like to live with?!—a developmental perspective on a personalized, life-long robot companion. In: ROMAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759) (2004) IEEE, pp 17–22
- Rossi S, Ferland F, Tapus A (2017) User profiling and behavioral adaptation for HRI: a survey. *Pattern Recogn Lett* 99:3–12. <https://doi.org/10.1016/j.patrec.2017.06.002>
- Ahmad M, Mubin O, Orlando J (2017) A systematic review of adaptivity in human–robot interaction. *Multimodal Technol Interact* 1(3):14
- Di Nuovo A, Varrasi S, Lucas A, Conti D, McNamara J, Soranzo A (2019) Assessment of cognitive skills via human–robot interaction and cloud computing. *J Bionic Eng* 16(3):526–539. <https://doi.org/10.1007/s42235-019-0043-2>
- McCambridge J, Witton J, Elbourne DR (2014) Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol* 67(3):247–253. <https://doi.org/10.1016/j.jclinepi.2013.08.015>
- Locke SD, Gilbert BO (1995) Method of psychological-assessment, self-disclosure, and experiential differences—a study of computer, questionnaire, and interview assessment formats. *J Soc Behav Pers* 10(1):255–263
- Paunonen SV (2003) Big five factors of personality and replicated predictions of behavior. *J Pers Soc Psychol* 84(2):411–424. <https://doi.org/10.1037/0022-3514.84.2.411>
- Paunonen SV, Ashton MC (2001) Big five factors and facets and the prediction of behavior. *J Pers Soc Psychol* 81(3):524–539. <https://doi.org/10.1037/0022-3514.81.3.524>
- Robert L (2018) Personality in the human robot interaction literature: a review and brief critique. In: Robert LP (eds) Personality in the human robot interaction literature: a review and brief critique, proceedings of the 24th Americas Conference on information systems, Aug 2018, pp 16–18
- Chien S-E, Chu L, Lee H-H, Yang C-C, Lin F-H, Yang P-L, Wang T-M, Yeh S-L (2019) Age difference in perceived ease of use, curiosity, and implicit negative attitude toward robots. *ACM Trans Hum Robot Interact* 8(2):1–19. <https://doi.org/10.1145/3311788>
- Edelstein BA, Woodhead EL, Segal DL, Heisel MJ, Bower EH, Lowery AJ, Stoner SA (2007) Older adult psychological assessment: current instrument status and related considerations. *Clin Gerontol* 31(3):1–35. <https://doi.org/10.1080/07317110802072108>
- Gosling SD, Rentfrow PJ, Swann WB (2003) A very brief measure of the big-five personality domains. *J Res Pers* 37(6):504–528. [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1)
- Lakens D (2017) Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Pers Sci* 8(4):355–362. <https://doi.org/10.1177/1948550617697177>
- Kennedy A (2000) Attention allocation in reading: sequential or parallel? In: Reading as a perceptual process. Elsevier, pp 193–220. <https://doi.org/10.1016/B978-008043642-5/50011-5>
- Fougnie D, Marois R (2011) What limits working memory capacity? Evidence for modality-specific sources to the simultaneous storage of visual and auditory arrays. *J Exp Psychol Learn* 37(6):1329–1341. <https://doi.org/10.1037/a0024834>
- Lucas GM, Gratch J, King A, Morency L-P (2014) It's only a computer: virtual humans increase willingness to disclose. *Comput Hum Behav* 37:94–100. <https://doi.org/10.1016/j.chb.2014.04.043>
- Iwasa H, Yoshida Y (2018) Psychometric evaluation of the Japanese version of Ten-Item Personality Inventory (TIPI-J) among middle-aged, and elderly adults: concurrent validity, internal con-

- sistency and test–retest reliability. *Cogent Psychol* 5(1):1426256. <https://doi.org/10.1080/23311908.2018.1426256>
30. Bopp KL, Verhaeghen P (2005) Aging and verbal memory span: a meta-analysis. *J Gerontol B Psychol* 60(5):223–233. <https://doi.org/10.1093/geronb/60.5.p223>
 31. Pollatsek A, Romoser MR, Fisher DL (2012) Identifying and remedial failures of selective attention in older drivers. *Curr Dir Psychol Sci* 21(1):3–7. <https://doi.org/10.1177/0963721411429459>
 32. Slawinski EB, Goddard KM (2001) Age-related changes in perception of tones within a stream of auditory stimuli: Auditory attentional blink. *Can Acoust* 29(1):3–12
 33. Levy B (1996) Improving memory in old age through implicit self-stereotyping. *J Pers Soc Psychol* 71(6):1092–1107. <https://doi.org/10.1037/0022-3514.71.6.1092>
 34. Armstrong B, Gallant SN, Li L, Patel K, Wong BI (2017) Stereotype threat effects on older adults' episodic and working memory: a meta-analysis. *Gerontologist* 57(suppl_2):S193–S205. <https://doi.org/10.1093/geront/gnx056>
 35. Deary IJ, Corley J, Gow AJ, Harris SE, Houlihan LM, Marioni RE, Penke L, Rafnsson SB, Starr JM (2009) Age-associated cognitive decline. *Br Med Bull* 92:135–152. <https://doi.org/10.1093/bmb/ldp033>
 36. Hedden T, Gabrieli JDE (2004) Insights into the ageing mind: A view from cognitive neuroscience. *Nat Rev Neurosci* 5(2):87–96. <https://doi.org/10.1038/nrn1323>
 37. Sliwinski MJ, Hofer SM, Hall C (2003) Correlated and coupled cognitive change in older adults with and without preclinical dementia. *Psychol Aging* 18(4):672–683. <https://doi.org/10.1037/0882-7974.18.4.672>
 38. Mroczek DK, Almeida DM (2004) The effect of daily stress, personality, and age on daily negative affect. *J Pers* 72(2):355–378. <https://doi.org/10.1111/j.0022-3506.2004.00265.x>
 39. Uchino BN, Berg CA, Smith TW, Pearce G, Skinner M (2006) Age-related differences in ambulatory blood pressure during daily stress: evidence for greater blood pressure reactivity with age. *Psychol Aging* 21(2):231–239. <https://doi.org/10.1037/0882-7974.21.2.231>
 40. Otoole BI, Stankov L (1992) Ultimate validity of psychological tests. *Pers Individ Differ* 13(6):699–716. [https://doi.org/10.1016/0191-8869\(92\)90241-G](https://doi.org/10.1016/0191-8869(92)90241-G)
 41. Yarkoni T, Westfall J (2017) Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci* 12(6):1100–1122. <https://doi.org/10.1177/1745691617693393>
 42. Calo MR (2011) 12 robots and privacy. In: *Robot ethics: the ethical and social implications of robotics*, p 187
 43. Pawlik K, Rosenzweig MR (2000) *The international handbook of psychology*. Sage, London
 44. Creech WA (1966) Psychological testing and constitutional rights. *Duke Law J* 2:332–371
 45. Krupp MM, Rueben M, Grimm CM, Smart WD (2017) A focus group study of privacy concerns about telepresence robots. In: 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN) IEEE, pp 1451–1458
 46. Sutanto J, Palme E, Tan CH, Phang CW (2013) Addressing the personalization-privacy paradox: an empirical assessment from a field experiment on smartphone users. *MIS Q* 37(4):1141–1141+. <https://doi.org/10.25300/Misq/2013/37.4.07>
 47. Bolarinwa OA (2015) Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Niger Postgrad Med J* 22(4):195–201. <https://doi.org/10.4103/1117-1936.173959>
 48. Drost EA (2011) Validity and reliability in social science research. *Educ Res Percept* 38(1):105–123
 49. Manning C, Raghavan P, Schütze H (2008) *Introduction to information retrieval*. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9780511809071>
 50. Li Q (2013) A novel Likert scale based on fuzzy sets theory. *Expert Syst Appl* 40(5):1609–1618. <https://doi.org/10.1016/j.eswa.2012.09.015>
 51. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, Surian D, Gallego B, Magrabi F, Lau AYS, Coiera E (2018) Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc* 25(9):1248–1258. <https://doi.org/10.1093/jamia/ocy072>
 52. Provoost S, Lau HM, Ruwaard J, Riper H (2017) Embodied conversational agents in clinical psychology: a scoping review. *J Med Internet Res* 19(5):e151. <https://doi.org/10.2196/jmir.6553>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Tsung-Ren Huang received his B.S. degree in mathematics from National Taiwan University, Taiwan, in 2001, M.S. degree in physics from National Taiwan University, Taiwan, in 2003, and Ph.D. degree in cognitive and neural systems from Boston University, USA, in 2009. He is currently an associate professor in psychology with National Taiwan University, Taiwan. His research interests include computational cognitive neuroscience, big data analytics, artificial intelligence, and robotics.

Yu-Wei Liu received his B.S. in psychology from National Taiwan University, Taiwan, in 2019. He is currently pursuing his M.S. degree in psychology at National Taiwan University, Taiwan.

Shin-Min Hsu received her B.S. degree in biomechanics engineering from National Taiwan University, Taiwan, in 2017. She is currently working toward her M.S. degree in psychology at National Taiwan University, Taiwan.

Joshua O. S. Goh received his bachelor's degree in psychology at the National University of Singapore, Singapore, in 2002, and Ph.D. degree in psychology at the University of Illinois at Urbana-Champaign, USA, in 2009. He was also a postdoctoral research associate at the National Institute on Aging, National Institute of Health, USA, where he worked with the Baltimore Longitudinal Study of Aging. He is now an associate professor at the Graduate Institute of Brain and Mind Sciences, College of Medicine, National Taiwan University. His main research interests include cognition, aging, culture, decision-making, and brain imaging.

Yu-Ling Chang received her B.S. and M.S. degrees in Psychology at the National Taiwan University, Taiwan, and Ph.D. degree in clinical psychology at the University of Florida, USA. She is now an associate professor at the Psychology Department, National Taiwan University. Her main research interests focus on cognitive aging, human memory and executive function, early detection of neurodegenerative diseases, and brain imaging.

Su-Ling Yeh received her B.S. and M.S. degrees in Psychology from National Taiwan University (NTU), Taiwan, and Ph.D. degree in cognitive psychology from the University of California, Berkeley, USA. Since 1994, she has been with the Department of Psychology, NTU and was awarded Lifetime Distinguished Professorship in 2012. She is a recipient of Academic award of Ministry of Education and Distinguished Research Award of National Science Council of Taiwan. She is an APS (American Psychological Science) fellow, and 2019-20 Stanford-Taiwan Social Science Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford University. She serves as associate director of NTU Center for Artificial Intelligence and Advanced Robotics and Editorial Board of Scientific Reports. Her research interests include cognitive neuroscience, perception, attention, consciousness, multisensory integration, aging, and applied research on display technology, eye tracking device, affective computing, and AI/robots.

Li-Chen Fu received B.S. degree from National Taiwan University, Taiwan, in 1981, and M.S. and Ph.D. degrees from University of California, Berkeley, U.S.A. in 1985 and 1987, respectively. Since 1987, he joined Dept. of Electrical Engineering and Dept. of Computer Science and Information Engineering, National Taiwan University (NTU), Taiwan, R.O.C. as a faculty member, and was awarded university's Lifetime Distinguished Professorship in 2007 and Chair Professorship in 2020. Currently, he serves as Director of NTU Center for Artificial Intelligence (AI) and Advanced Robotics as well as Co-director of MOST (Ministry of Science and Technology)/NTU Joint Research Center for AI Technology and All Vista Healthcare. So far, he has received numerous recognitions, including IEEE Fellow (2004), IFAC Fellow (2017), and National Chair Professorship of Taiwan (2019). Internationally, he is serving as Editor-in-Chief of the Asian Journal of Control, and Advisory Committee member of Asian Control Association. His research interests include social robotics, smart home, visual detection and tracking, virtual reality, and control theory & applications.